# Causal Discovery Methods for Climate Networks

**Imme Ebert-Uphoff***
School of Mechanical Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0405
ebert@me.gatech.edu

**Yi Deng**
School of Earth and Atmospheric Sciences
Georgia Institute of Technology
Atlanta, GA 30332-0340
yi.deng@eas.gatech.edu

**Research Report GT-ME-2010-001**
School of Mechanical Engineering
Georgia Institute of Technology

December 20, 2010

**Abstract**

This paper suggests new methods for the development of network models in climate research. Current climate networks, first introduced in 2004 by Tsonis and Roebber, define network edges based on correlation of node pairs, resulting in a correlation network. The key idea of this paper is to introduce techniques from causal reasoning to derive climate networks, specifically constraint based structure learning. This approach is expected to yield networks that better represent the causal connections in the network, by containing less edges and with all causal pathways still present. The anticipated advantage of a network with less edges is a more manageable model size that makes it easier to gain new insights about causal relationships in the climate system.

The goal of this paper is to provide researchers in the climate area with an intuitive understanding of the causal discovery process, specifically of constraint based structure learning. We review the basic principles of constraint based structure learning, namely how cause-and-effect relationships of variables can be learned from observational data using *conditional* independence tests. Tutorial-style examples illustrate this process. Finally, we review available algorithms and software packages from other disciplines that can be applied to derive climate networks.

There are *no* simulation results provided in this paper (work in progress), thus we do not yet know how much reduction is achieved through this method compared to existing methods. However, applications of similar techniques for protein interaction modeling has yielded tremendous savings, making it possible to gain significant understanding of causal pathways from the obtained network graphs.

*Keywords:* Climate network, causal discovery, structure learning, constraint based learning, conditional independence, mutual information, probabilistic graphical model.

*Joint Appointment with the Robotics and Intelligent Machines Center, School of Interactive Computing, College of Computing, Atlanta, GA 30308.

# 1 Motivation

In their seminal papers Tsonis and Roebber [30] and Tsonis et al. [33] introduced tools from network analysis to the field of climate research. They described how atmospheric fields - or other quantities - can be used to define a network of nodes, where each node represents a point on a global grid. The network contains an edge between any pair of nodes if the cross correlation of their atmospheric fields exceeds a certain threshold, $\|r\| > r^*$. Due to the way the edges are defined we refer to this type of network as *correlation network*. Since correlation networks were introduced to climate science in 2004 [30], there has been a flurry of research activity in this area. A summary of the progress, opportunities and challenges of networks in climate science was presented in 2010 by Steinhaeuser et al. [29].

## 1.1 Properties of Climate Networks

Once a network is obtained its global and local properties can be evaluated. Local properties include the number of connections per node, which are used to identify clusters and isolated nodes [33]. Steinhaeuser et al. [28] developed an algorithm for *community detection* that they apply to correlation networks to discover clusters corresponding to climate regions.

Global properties include the average path length between nodes and clustering coefficients. The network topology is often summarized by categorizing the network as one of four basic types, regular network, classic random network, small-world network or network with a given degree distribution (Tsonis et al.[33]). For example, a climate network based on atmospheric fields was shown to result in a small-world network [33].

Donges et al. [12] recently introduced a very interesting local measure for climate networks, the *betweenness centrality* of a network node. Betweenness centrality (BC) of a node measures whether the node is traversed by a large number of all existing shortest paths in the network. In the context of climate networks BC can be interpreted as a local measure of energy transport in the network. Using this measure they identify peculiar wave-like structures with high BC values in climate networks, which they call the *backbone of the climate network*. They conclude that the backbone represents pathways of global energy and dynamical information flow in the climate system.

Furthermore, networks present an excellent tool for visualization that can be used to explore whether observed/suspected teleconnections can be explained by pathways in the network. For example, by calculating separate networks over a sequence of time periods a temporal sequence of networks is obtained. By analyzing changes in topology, patterns of climate changes can be detected. Network edges can be categorized into *stable edges* that persist over a long period of time, and *blinking edges* (generally called *blinking links* in literature) that go on and off over time [16, 36]. Using this type of analysis Tsonis et al. [31, 32] and Gozolchiani et al. [16, 35] were able to relate global network changes to El Niño activity (there are less edges during El Niño). Clearly, much has already been learned from climate networks in the few years of their existence.

## 1.2 Other Ways to Define Climate Networks

To date climate networks are almost always defined as correlation networks. However, two other definitions have recently been proposed, *MI networks* and *phase synchronization networks*, which are discussed below.

Donges et al. [12] define *MI networks*, where *mutual information (MI)* between any two nodes is used as a measure to detect the edges. They find that the resulting MI networks are largely similar to the corresponding correlation networks, but that MI networks are better at detecting edges corresponding to *nonlinear* statistical interrelationships, a finding that matches our discussion of mutual information in Section 4.3. Note that both correlation networks and MI networks *decide whether an edge exists between two nodes in the network based only on a test involving those two nodes*. However, considering only the data for those two nodes it is impossible to distinguish between nodes that are connected directly and those that are connected only through intermediate variables. Therefore these networks result in unnecessarily many edges. Obviously, the more edges there are, the harder it is to gain information from the network and visualization may become pointless. Thus one may look for meaningful ways to reduce the number of edges in the network. The standard approach is to raise the correlation threshold, $r^*$, thus requiring connections to be stronger in order for them to show up as an edge in the network [29]. However, much relevant information may be lost through that process.

Yamasaki et al. [36] seek additional clues about relationships in the data using temporal signatures. They define the *phase synchronization network*, where pairs of nodes are connected by an edge based on their *synchronization strength*. Synchronization strength is a concept from time series analysis and it measures the coupling of two cyclic signals. While edges are also defined only by pair-wise tests for these networks, analyzing their temporal relationship may yield additional information about causal relationship. It turns out that for their example the resulting phase synchronization networks are very similar to the corresponding correlation networks in most geographic areas with some local differences in the remaining areas. Those differences are not yet well understood. Nevertheless, synchronization strength is an interesting concept and phase synchronization networks may complement other approaches in the future.

### 1.3 Climate Networks through Causal Discovery

We propose a more direct approach to finding causal relationships from data - and thus deriving networks with less edges - using methods from causal discovery. Causal discovery, a field established by Pearl [22] in the 1980s, seeks to learn as much as possible about causal connections in a system from its data. Since it is well known that correlation of two variables does *not* imply causation, tests other than cross correlation must be used to identify causal relationships. The basis of causal discovery is to use - in addition to the common independence tests that only involve two variables - also *conditional* independence tests that involve three or more variables. They are discussed in detail in Section 4.3.

Causal discovery, specifically constraint based structure learning, uses conditional independence tests of statistical data to infer as much as possible about causal connections in the system, and describes the results in the form of graphs. A graph obtained through structure learning provides an alternative network description, which we refer to as *causal discovery network*. The properties of the causal discovery network can then be analyzed in the same way as correlation networks (Section 1.1), but causal discovery networks are expected to generally contain less edges, sometimes significantly less.

Causal discovery has been used this way in other disciplines. There are several areas dealing with large networks, namely social networks, text mining and computational biology. Out of those computational biology has been the most active research area for causal discovery methods in recent years. Researchers in computational biology train networks to identify protein/gene interaction in cells based on expression data [20, 15]. Customized causal discovery methods have been developed in this area for networks containing tens of thousands of nodes [20]. Those algorithms may provide a powerful alternative for the development of climate networks with a (nearly) minimal number of edges.

## 2 Organization

The remainder of this paper is organized as follows. Section 3 introduces some terminology from graphs and probabilistic graphical models. Section 4 discusses why causality is an important concept for the derivation of climate networks and discusses the basics of causal reasoning, especially conditional independence tests. Section 5 describes how conditional independence tests can be used for constraint based structure learning. Section 6 discusses a variety of algorithms available for structure learning that may be suitable for climate networks. Section 7 discusses some additional ideas such as incorporating external influences as nodes in causal discovery networks and using varying resolution. Section 8 presents conclusions and future work.

## 3 Notation

Graphs are a convenient way to represent conditional independencies between random variables. They are a powerful tool to visualize dependencies between variables in the system. They are also a convenient computational structure that encodes the dependencies in a compact way for use in a great variety of computational algorithms. This section introduces a few key concepts for graphs.

A **graph** $G = (V, E)$ consists of a set of vertices, V, and a set of edges that connect pairs of vertices. **Directed** graphs have a unique direction assigned to each of the edges, while **undirected** graphs have no direction assigned to any of the edges. A directed graph is **acyclic** if it does not contain cycles, i.e. starting at

any node and following the arrow directions one can never get back to the start node. A **chain graph** allows a mixture of directed and undirected edges, but those are not considered further in this paper.

The vertices of a graph are often called **nodes**. The set of nodes that share an edge with node $X$ in a graph are called the **neighbors** of $X$. In an undirected graph one only speaks of neighbors. In a directed graph one distinguishes between child and parent nodes. If $X$ and $Y$ are neighbors in a directed graph and the arrow points from $X$ to $Y$, then $X$ is called a **parent** of $Y$ and $Y$ is called a **child** of $X$.

**Probabilistic graphical models** combine tools from graph theory with probability theory. Graphical models are a popular modeling tool for systems with uncertainty. The most common type is the **Bayesian Network**, also known as Bayes Net or Belief Network. A Bayesian Network model consists of a directed acyclic graph and a probability distribution assigned to each node which defines the probability of the node's state based on the states of its parents. Bayesian networks have found widespread application in many disciplines, from medical diagnosis [4] to protein interaction [20] and have recently emerged in selected applications of atmospheric sciences, such as precipitation modeling [9, 5, 19], forecasts of severe weather [1] and air pollution modeling [11].

The **Markov Network**, also known as Markov Random Field, is based on an undirected graph. A Markov network can represent certain dependencies that a Bayesian network cannot (such as bi-directional and cyclic dependencies); on the other hand, it cannot represent certain dependencies that a Bayesian network can (such as v-structures, see Section 5.3) [18].

Within the scope of this paper we do not deal explicitly with any of the probabilities. All we care about are the structure learning algorithms that were developed for graphical models to learn the structure of the underlying graphs.

# 4 Causal Reasoning

In many applications probabilistic graphical models are used to derive causal models. For example, in a Bayesian network the arrows of the directed graph can often be interpreted as going from *cause* to *effect*. In a Markov model the edges of the graph are undirected so causal influences can go in both directions.

## 4.1 Using Causality to Achieve Minimality

Philosophers and mathematicians have struggled over millennia to derive a concise definition of *causality*. The discussion was brought to a solid footing in the late 1980s through the introduction of *Causal Calculus* by Rebane and Pearl [24], and the subsequent development of the first algorithm for the recovery of cause effect relationships from statistical data (Pearl [22]).

Pearl [22] writes: *What are the merits of these fictitious variables called causes that make them worthy of such relentless human pursuit, and what makes causal explanations so pleasing and comforting once they are found? We take the position that human obsession with causation, like many other psychological compulsions, is computationally motivated. Causal models are attractive mainly because they provide effective data structures for representing empirical knowledge [...].*

Indeed it turns out for Bayesian networks that if the edges in the directed graph are based completely on causal relationships that this Bayesian Network is guaranteed to provide the most compact way of representing the system's joint probability. In other words, the underlying graph requires the least number of edges and the associated probability tables require the least number of probabilities to define the full model. The causal model is the *minimal* model. Similar statements hold for Markov models. This fact is very important for the application of climate networks, because our goal is precisely to find the network representation with the minimal number of edges.

## 4.2 The Match Example

To illustrate several concepts from causal reasoning we introduce the match example in this section. One can light a match by striking its head on sand paper. The friction between sand paper and match head causes heat, which in turn starts a chemical reaction in the match head, setting the match on fire. This process can be described by three variables:

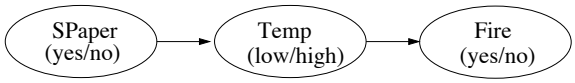


Figure 1: Intuitive causal graph for match example

- $SPaper$ (yes/no), which indicates whether the match head recently touched the sand paper,
- $Temp$ (low/high), which indicates the temperature of the match head, and
- $Fire$ (yes/no), which indicates whether the match was set on fire.

By understanding the physical processes involved we can describe the causal connections intuitively in the graphical form shown in Figure 1.

Note that Figure 1 shows arrows from $SPaper$ to $Temp$ and from $Temp$ to $Fire$. However, there is *no* edge between $SPaper$ and $Fire$. In fact the cause-effect relationship between $SPaper$ and $Fire$ *always* goes through the variable $Temp$. In other words, if we want to make a prediction for whether the match is on fire, and we already know the temperature of the match head, we really do not gain any additional information by knowing whether the match recently touched the sand paper. In essence the variable $Temp$ blocks the information flow from $SPaper$ to $Fire$. In probabilistic terms we say that random variable $Fire$ is *conditionally independent* from $SPaper$ given $Temp$.

## 4.3 Independence and Conditional Independence

Causal discovery, specifically constraint based structure learning, is solely based on tests for independence and conditional independence. Thus this section is dedicated to reviewing these concepts and potential tests for them. For ease of notation we define these concepts here for discrete random variables, but all of the definitions extend to continuous variables.

Two discrete random variables, $X$ and $Y$, are said to be *independent* of each other if

$$\mathrm{P}(X{=}x, Y{=}y) \;=\; \mathrm{P}(X{=}x) \cdot \mathrm{P}(Y{=}y).$$

for any $x$, $y$. In other words two random variables are independent if knowing the state of one does not tell you anything about the state of the other.

Two discrete random variables, $X$ and $Y$, are *conditionally independent* given a third random variable, $Z$, if for each value of $Z$, the variables $X$ and $Y$ are independent of each other. Denoting as $\mathrm{P}(X = x|Y = y)$ the conditional probability that $X$ takes the state $x$, conditioned on the fact that $Y$ is in state $y$, this conditional independence can be expressed as

$$\mathrm{P}(X = x|Y = y, Z = z) \;=\; \mathrm{P}(X = x|Z = z)$$

for any $x$, $y$ and $z$ with $P(Z = z) > 0$. Thus if $X$ and $Y$ are conditionally independent given $Z$, then if you are interested in the state of $X$ and already know the state of $Z$, then knowing $Y$ in addition does not add *any* new information. One can say that $Z$ blocks the information flow from $X$ to $Y$.

We saw an example of a conditional independence relationship in the match example above. In the example the conditional independence was concluded from our understanding of the physical problem. However, in structure learning we want to learn unknown conditional independencies in a system based on data. For that we need tests for independence and conditional independence (CI).

There is a great variety of measures that can be used to test for independence and conditional independence. In statistics the traditional choice is cross correlation as measure for independence and partial correlation as measure for conditional independence. If all considered variables are multivariate Gaussian, then the well-known partial correlation coefficient, $\rho_{XY|Z}$, is zero if and only if X is conditionally independent from Y given Z. Thus partial correlation is a good test in that case. However, for general distributions the above relationship does not hold. Nevertheless, partial correlation is sometimes used in practice, because it is well known and thus convenient to use.

In information theory the most common measure of independence is *mutual information*. Mutual information is based on the concept of entropy, $U(X)$, which measures the amount of uncertainty contained in a variable, $X$. For two discrete random variables, $X$ and $Y$, mutual information can be defined as

$$MI(X,Y) = U(Y) - U(Y|X) = \sum_{x,y} P(X=x, Y=y) \log_2 \frac{P(X=x, Y=y)}{P(X=x) \cdot P(Y=y)}.$$

*Conditional mutual information* is an extension of mutual information suitable to measure conditional independence. For three discrete random variables $X$, $Y$ and $Z$, conditional mutual information can be defined as follows

$$MI(X,Y|Z) = \sum_{x,z} P(X=x, Z=z) \sum_{y} P(Y=y|X=x, Z=z) \log_2 \frac{P(Y=y|X=x, Z=z)}{P(Y=y|Z=z)}.$$

(The definition of conditional mutual information for continuous variables can be found in [3].) Conditional mutual information compares the uncertainty in $X$ if we know the state of $Z$ to the uncertainty in $X$ if we know the states of both $Z$ and $Y$. If the amount of uncertainty is identical, then MI vanishes and $X$ is considered conditionally independent for $Y$ given $Z$.

In contrast to partial correlation, conditional mutual information is a good CI test for *any* type of distributions. This matches the observations of Donges et al. [12] who found that independence tests based on correlation only detect linear relationships, while mutual information also detects non-linear relationships and concluded that the highly nonlinear processes at work in the climate system calls for the application of nonlinear methods, such as mutual information, to obtain more reliable results.

Vanishing conditional mutual information is the most common CI test used in structure learning, but many other measures exist. For an extensive review see [2].

Note that the definition of conditional independence as well as of the actual CI tests apply not only if $Z$ represents a single random variable, but just as well for a *set* of several random variables, $Z = \{Z_1, \ldots, Z_k\}$.

Unfortunately, in practical use CI tests face some real limitations:

1. Even if two variables are perfectly conditionally independent in theory, due to noise in the statistical data CI test result will rarely come out to be exactly zero. Thus all CI tests are used in combination with a threshold that determines when variables are considered to be independent.

2. The reliability of the CI test depends on the sample size. The more samples are available the more reliable the result.

3. Reliability declines with increasing number, $k$, of conditioning variables, $Z_1, \ldots, Z_k$, so we should always try to avoid large conditioning sets.

# 5 Structure Learning Through CI Tests

There are two primary methods for structure learning. One is a score based method that learns the graphs along with probabilities and uses some type of optimization routine to maximize the fit of the model. The most popular algorithm is the K2 algorithm [10] which yields good results to learn the structure and probabilities for Bayesian networks of small to medium size. Considering that we may deal with hundreds, thousands or tens of thousands of nodes in a climate network - and thus up to millions (!!) of probability parameters -, addressing the full optimization problem seems infeasible. Friedman et al. [14] proposed a clever way to limit the search space that makes it feasible for larger networks. Their approach is briefly discussed in Section 6.

However, in this paper we focus on the second method, constraint based learning, because it is generally better suited to deal with large numbers of nodes. Constraint based learning breaks the learning process of a graphical model up into two steps. First CI tests are used to learn as much as possible about the structure of the underlying graph. Once a graph structure is established the probability parameters are learned in the second step. Since we only care about the graph structure anyway we can simply stop the learning process after the first step and thus never deal with any probability parameters.

For most climate networks undirected graphs may the best choice because we cannot exclude the possibilities of bi-directional cause and effect relationships, i.e. a pair of nodes may interact with each other in

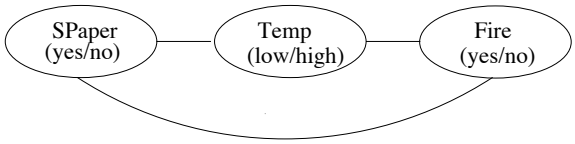


Figure 2: Correlation network for match example

both directions. Thus constraint based structure learning of *undirected* graphs is the method we are most interested in. However, we include structure learning of *directed* graphs for several reasons. (1) The learning of directed graphs and undirected graphs is interwoven and there is much overlap. (2) More algorithms have been developed for directed graphs; (3) Directed graphs can be converted to the corresponding undirected graphs. (4) Directed graphs also have applications in a variety of other problems in atmospheric sciences and may become more prevalent in the future.

Note that by *causal discovery networks* we mean both undirected and directed graphs obtained through causal discovery methods.

## 5.1 Footprints of causal relationships in data

To recover causal relationships from data we need to learn to read their footprints, i.e. the traces they leave in the data. There are two main concepts to understand

1. The difference between direct and indirect connections;
2. So called *V-structures*

Section 5.2 illustrates the first of these concepts, Section 5.3 illustrates the second.

## 5.2 Testing for Direct Connections

To understand how structure learning with CI tests may work - and why! - we revisit the match example. For the moment let us forget everything we know about the physical mechanisms in the match example. Instead we are given statistical data obtained by observing the variables over an extended amount of time. There is some uncertainty in the system. For example heat may be generated occasionally through other causes, e.g. by someone holding the match close to another flame once in a while, or the friction on the sand paper may not be sufficient to start the flame. We now have a large data base of observed cases, where each case lists the state of all three random variables. Our task is to learn as much as possible about causal connections for this example from the data.

First we try the correlation network. The data would reveal $SPaper$ to be closely correlated with $Temp$, and $Temp$ to be closely correlated to $Fire$. As a result $SPaper$ is also closely related to $Fire$, resulting in the correlation network in Figure 2, where all nodes are connected to each other and none of the arrows have a direction associated with them.

Now let us apply CI tests. Since this example only has three nodes only three CI tests would have to be performed. Namely, we would test whether any two of the variables are conditionally independent given the third variable. For large enough sample size only one CI test would come back negative, namely only $SPaper$ and $Fire$ are conditionally independent given $Temp$. This makes intuitive sense, because if we want to know whether the match is likely on fire, and we already know the temperature of the match is low/high, it really does not matter whether the match recently touched the sand paper. One can say that the intermediate variable, $Temp$, blocks the flow of information from $SPaper$ to $Fire$.

Based on that CI test result we can now eliminate the edge between $SPaper$ and $Fire$ and obtain the undirected graph in Figure 3.

If we wanted to learn a directed graph from that information there are actually three possible graphs based on the results of the conditional independence tests. Those three graphs are shown in Figure 4. On the top is the correct graph, identical to the one we intuitively came up with in Figure 1. The other two vary in the direction of at least one edge. Just based on data it is actually not possible to determine which of the three

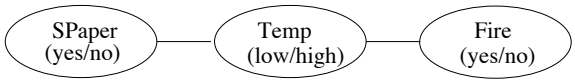


Figure 3: Undirected causal discovery graph for match example

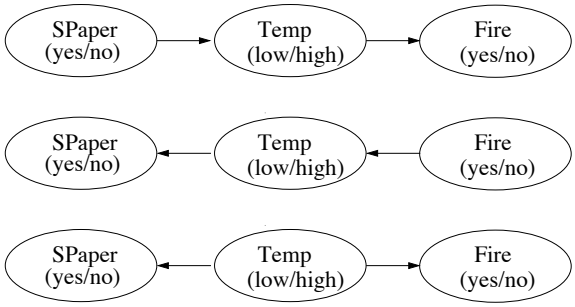


Figure 4: Three different directed causal discovery graphs for match example

graphs in Figure 4 is correct. The three graphs are indistinguishable from a structure learning perspective. One says they are *Markov equivalent*. Note that the graph with both arrows pointing toward $Temp$ is not included in Figure 4. That graph is actually eliminated because the data does not show a *v-structure*, as explained in Section 5.3.

The match example is very simple, but it demonstrates a basic principle of how causal reasoning can be used to eliminate one or more edges from a graph.

## 5.3 Finding Edge Directions through V-Structures

A *v-structure* in a directed graph is a child node that has at least two parents which are not connected to each other. v-structures, also known as *unshielded colliders*, play a key role in causal reasoning because they are the key indicators for the *direction* of causal relationships. The following application provides an example of a v-structure.

Whether a person develops lung cancer depends among other things on age and smoking habits. In other words the variables $Age$ and $Smoking$ are causes (parents) of the effect (child) $LungCancer$. Furthermore, let us say that for the considered population the age of a person does not significantly impact whether he/she smokes or not. Thus $Age$ and $Smoking$ are considered independent of each other and the intuitive causal graph shown on the left of Figure 5 does not show an edge between them.

This causal graph contains a v-structure at $LungCancer$, since this node has two parents which are not connected to each other. The name v-structure comes from the fact that these three nodes form the shape of a "V" if we follow the convention of placing causes higher up on the page than effects.

V-structures leave a distinct footprint that can be detected in the corresponding data, and thus used to determine directions in a directed graph representation. Namely, the parent nodes are *independent* of each other, but they become *conditionally dependent* if the state of the child is known. Let us illustrate this conditional dependency using the lung cancer example. We made the assumption that $Age$ is independent of $Smoking$, i.e. knowing the age of a person does not tell me anything about his/her smoking habits. However, if we know the status of the variable $LungCancer$, say that a person has been diagnosed with lung cancer, then the parent nodes become dependent. For example, knowing that a person with lung cancer diagnosis is of young age raises the probability that the person is smoking, because lung cancer patients often have at least one of the two major risk factors, increased age or smoking.

v-structures thus leave a very particular signature in the data. Namely, they consist of two nodes that are independent of each other, but they are dependent on a third and the two nodes become conditionally dependent *given* the state of the third. Identification of v-structures is used in the structure learning of di-

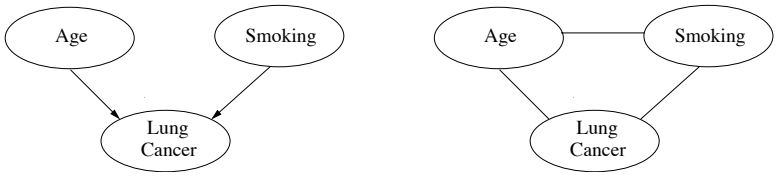


Figure 5: Intuitive causal graph (left) and correlation network (right) for lung cancer example

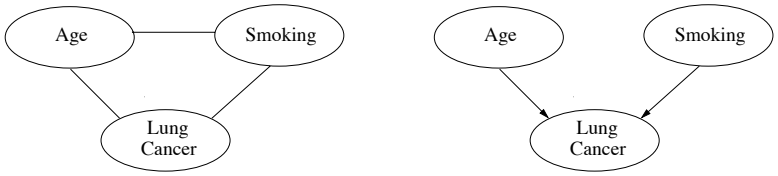


Figure 6: Undirected (left) and directed (right) causal discovery graph for lung cancer example

rected graphs to identify the direction of as many arrows as possible. Other arrow directions are determined afterward based on the constraints that (1) no directed cycles are allowed; (2) no additional v-structures can be created.

v-structures also play a special role for undirected graphs. An undirected graph is unable to represent the independence relationships of a v-structure, resulting in an additional edge between the parents. See Section 5.5 for conversion from directed to undirected graphs.

Figure 5 on the right shows the correlation graph for this example. Figure 6 shows on the left the undirected and on the right directed graph that would be obtained through structure learning. Because of the v-structure learning of the directed graph in this example yields one unique graph and this graph perfectly matches the intuitive causal graph. (In contrast three different directed graphs were obtained for the match example in Figure 4, because of the lack of a v-structure in that application.)

Because of the v-structure the learned undirected graph contains one more edge than the learned directed graph, namely between the parent nodes $Age$ and $Smoking$. While in this particular example the correlation graph is identical to the learned undirected graph, in general the undirected graph is expected to have significantly less edges than the corresponding correlation graph.

## 5.4 Equivalence classes of Directed Graphs

Causal discovery from observed data is only able to determine directed graphs up to an equivalence class, namely the set of Markov equivalent graphs [21]. This equivalence class may contain one or more graphs. Only an intervention analysis - where we actively manipulate the states of some variables in targeted experiments - can reveal additional causal relationships, see Pearl [23] or Murphy [21].

Two directed graphs are called *Markov equivalent* if they represent the same set of independence relationships. As it turns out this equivalence can also be expressed as follows. Two directed graphs are Markov equivalent if they have the same set of edges (ignoring the edge direction) and the same set of v-structures. For example the three directed graphs in Figure 4 form a Markov equivalence class and it is not possible to further narrow down which graph is the correct one without performing intervention experiments.

## 5.5 Conversion from Directed to Undirected Graph

A directed graph can always be converted to its corresponding undirected graph by a process called moralization. That means that we first find all v-structures and *marry* all unmarried parents, i.e. we add an edge to any pair of parents with a joint child that do not already share an edge. Afterward all remaining arrow directions are dropped in the graph. Thus an undirected graph either has exactly as many or up to a few more edges as the corresponding directed graph.

It is easy to show that all directed graphs of a Markov equivalence class correspond to the same undirected graph. Thus structure learning may result in several directed graphs, but only one undirected graph.

| Algorithm | Package | Web location and comments |
|---|---|---|
| PC algorithm | TETRAD | http://www.phil.cmu.edu/projects/tetrad/tetrad4.html<br>Comments: One of the very first algorithms, still popular |
| Borgelt's algorithm | INeS | http://www.borgelt.net/ines.html<br>Comments: Undirected graphs |
| Protein interaction | ARACNE | http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE<br>Comments: Proposed for tens of thousands of nodes |
| PC, FCI, IC, IC$^*$, K2, Hill Climbing and more | Bayes Net Toolbox | http://code.google.com/p/bnt/<br>Comments: Matlab implementation of many algorithms |

Table 1: A sample of available software packages for structure learning (not an exhaustive list)

# 6 Available Structure Learning Algorithms

A large number of constraint based algorithms has been developed in the two decades since causal calculus was first introduced and software code for many of the these algorithms is readily available (Table 1). A review of all available structure learning algorithms is beyond this paper, we can only provide highlights. The purpose of this section is to help the reader navigate and select potential algorithms for climate network applications. For now our goal is to obtain causal discovery networks in the form of *undirected* graphs to represent climate interactions.

All existing network approaches for climate research, - correlation networks, MI networks and phase synchronization networks -, share the fact that whether two nodes are connected by an edge depends only on tests involving those two nodes. Thus the computational complexity of those approaches is always $O(N^2)$, where $N$ is the number of nodes in the network. In contrast, constrained based structure learning, seeking to include (direct) causal pathways only, needs to use conditional independence tests that may involve several additional nodes in order to decide whether an edge should be present between a node pair. Thus these algorithms are always of higher computational complexity than the existing methods. In view of the fact that climate networks may have thousands or even tens of thousands of nodes the key challenge is to find algorithms that can deal efficiently with such a large number of nodes.

Standard algorithms for constraint-based structure leaning use a series of independence tests and conditional independence tests on the data to determine the set of edges and the v-structures. From that information the corresponding equivalence class of directed graphs - and the moralized undirected graph - can be derived. The art of designing such an algorithm lies in choosing the order of all the tests to (1) reduce computational complexity and (2) increase robustness, since an error in one edge early on may cause a chain reaction of other edge errors later on. We also discuss algorithms targeted for networks with many nodes that use simplifying assumptions (short-cuts) to implement a trade-off between complexity and accuracy of the resulting graphs.

Due to the great popularity of Bayesian networks the majority of algorithms has been developed for directed graphs. The first algorithms proposed for causal discovery are the IC algorithm developed by Verma and Pearl [34, 23], and the SGS algorithm, developed by Spirtes, Glymour, Scheines [26, 27]. The SGS algorithm is exponential in the number of nodes and not very efficient. Spirtes et al. thus developed the PC algorithm [25, 27] to recover sparse networks, i.e. networks with few edges. The PC algorithm limits the number of edges allowed for each node to a fixed number, $k$, and the PC algorithm is exponential in $k$ only, rather than in the number of nodes. The PC algorithm has been used in practice to recover sparse networks with over a hundred nodes. The PC and IC algorithms both assume that all nodes of importance for causal relationships are included in the network. If, however, there are external variables that greatly influence the causal relationships, then so-called *hidden* nodes can be introduced to model those. The FCI algorithm is an extension of the PC algorithm and the IC$^*$ algorithm is an extension of the IC algorithm for that purpose.

Another popular approach is the algorithm by Cheng et al. [7] and Cheng et al [6]. The complexity of this algorithm is of polynomial order, namely or order $N^4$, where $N$ is the number of nodes. This reduction is based on the so-called monotone faithfulness assumption, which according to Chickering and Meek [8] is a bad assumption. Nevertheless, the algorithm is widely used in practice and appears to give good results.

When interested in undirected graphs, there is some unnecessary overhead involved in deriving all the directed graphs first and converting them to undirected graphs later. Borgelt [2] describes a modification of the structure learning algorithm by Cheng et al. [6] that calculates the undirected graphs directly, thus improving efficiency.

The biggest challenge for the novel application of these algorithms to model climate interactions is the large number of nodes required by climate models. Depending on the resolution of the grid there can be hundreds, thousands or tens of thousands of nodes. Some of the above algorithms may not scale up to so many nodes. Research in social networks, text mining (e.g. automatic text categorization [17]) and computational biology (identifying protein/gene interaction in cells based on expression data [20, 15]) all deal with large networks and we can learn from their approaches. Out of the above, computational biology appears to be the most active field to date to develop customized structure learning algorithms for networks with a large number of nodes. Margolin et al. [20] developed a successful approach for identification of interactions in gene regulatory networks. They replace CI tests by the use of the data processing inequality (DPI) which strictly speaking should only be applied to recover tree-structured graphs. However, through additional steps they are able to apply this algorithm also for other graphs. This algorithm is of lower complexity than the previous ones, $O(N^3)$, and the authors claim that it can be applied to recover networks with tens of thousands of genes [20]. The implementation of their algorithm is known as the ARACNE package. This algorithm is certainly a prime candidate to try for climate interaction networks.

Friedman et al. [15] use a score-based learning approach. In contrast to constraint based learning, score-based approaches formulate the search for graph candidates as an optimization problem. The problem with most such approaches is that the search space is super-exponential in the number of nodes [15] and thus not suitable for large networks. Friedman et al. therefore developed the *Sparse Candidate* algorithm [14] that identifies for each variable only a small number of candidate parents, thus greatly restricting the search space.

Zeng and Poh [38] and Zeng and Hernandez [37] propose a divide and conquer strategy to structure learning by learning local components first and joining the components to the complete network. In [37] the algorithm is tested for networks with up to 223 nodes and appears suitable for larger networks. One benefit of the algorithm is that it also performs well for smaller sample sets. There may also be value in applying this type of algorithms to recover local graph components of specific geographic areas, since the learned components may represent local knowledge more precisely in comparison to the full Bayesian networks when working with a small amount of data [39].

In conclusion, considering that CI tests are not perfect in the first place, and that we cannot expect to get a perfect model anyway, the final selection of an algorithm can only come from testing different algorithms on real climate data.

# 7 Choosing Network Nodes - Including External Influences and Other Ideas

Causal discovery theory is based on information theory and interprets *flow* in a climate network simply as transport of *information*. However, in a physical system it is worthwhile to contemplate by which physical means this information is being transported. Yamasaki et al. [36] state: *Information flows between [the nodes] in the form of heat transfer, wind, flow of water and transfer of other materials (represented by links). There are also global external changes which influence all the nodes in the climate network (for example – variation of the radiation from the sun).*

In the context of causal discovery, external variables such as the radiation of the sun, are called *hidden causes* or *latent variables* and there are methods for discovering those (see [23, 27] and the FCI and IC $^*$ algorithm mentioned in Section 6). In general, a hidden cause is a variable that has a causal effect on at least two nodes in the network, but that itself is not included in the network, thus leading to an incomplete - and sometimes misleading - model of causal relationships. Obviously, it is impossible to include all external variables in a climate model, otherwise the model would be infinitely complex. However, the influence of *some* external variable(s) may be so strong that a better model is achieved by including it explicitly in the network. In fact the number of edges in a causal discovery network may actually be *reduced* by including such hidden causes, because direct cause-effect relationships in the extended graph may replace many indirect pathways over long distances in the original graph. (In contrast, in a correlation network, including an external

variable would not change *any* of the existing edges, but only add new ones!) For example, we believe that adding an El Niño index as node to a climate network obtained through structure learning may yield very interesting results.

Fortunately, causal discovery methods offer immense flexibility in the modeling process. In fact, since they are based on the general concept of information flow, *any* type of variable can be added as a node to the network. Thus it is possible to add variables representing any type of physical quantity, as well as abstract variables that only have a few different states, as nodes to a climate network. (A good example of the seamless integration of abstract nodes and meteorological data nodes in a Bayesian network is provided by the Hailfinder project for severe weather forecasting [1].) A domain expert would be the best source of information on which external variable(s) to add to a network and the structure learning methods could then be applied to the enlarged set of nodes.

Several research groups analyzed temporal sequences of climate (correlation) networks and found that their topology changes over time and were able to track some of the changes to El Niño activity ([16, 36, 31, 32]). Yamasaki et al. [36] also found that *Between most of the pairs of nodes in the climate network there exists an indirect coupling. This means that a few localized severe events (such as massive heat transfer between the ocean and the atmosphere in a restricted zone of the pacific) can, in principle be felt as a change of the coupling between two nodes outside this zone. These changes may be tracked by rapid change in the correlation pattern between the two nodes.* All these observed changes in network topology indicate that some causal relationships are omitted in the network model, most likely due to hidden causes such as El Niño - another indication that adding such external variables is essential. In fact, our ideal goal is to include *all* causal relationships in our network model, including those that result in climate change. That would result in a stationary network model, i.e. a network whose topology does not change over time. In fact how stationary the climate network is can be seen as test whether all significant causes and all signficant causal relationships are included. To evaluate *how* stationary a network is one should not only consider the existence of edges, but also the *strength* of the edge connections, using measures such as the ones discussed in [13], but adjusted to undirected graphs [1]. This is important because a blinking edge of very low strength may be just be due to noise, while a blinking edge of high strength indicates a true change in the model. Once a stationary model is achieved it gives us some confidence that *all* causal relationships are included, even those that are responsible for climate change. It will be interesting to see whether including El Niño alone will result in a stationary network topology, or whether other external influences have to be added as nodes.

Another idea we plan to pursue is to define climate networks with varying resolution in different geographical areas. Resolution could be based on the similarity of activity of the individual nodes in those areas – the higher the similarity, the lower the resolution. A similarity index has yet to be defined. Another option is to use the BC measure of local energy transport introduced by Donges et al. [12] as criterion to decide on local resolution.

In an abstract way the ideas discussed in this section, combining similar nodes to a single node and including external variables as nodes, present a deviation from a *strictly geographical* selection of network nodes to node selection that *best represents causal relationships* of physical effects in the geographic area. In fact one can move freely anywhere between the two extremes, from purely geographical nodes to completely abstract nodes, depending on which aspects of climate one wants to focus on.

# 8 Conclusions and Future Work

In this paper we described why causal reasoning may play an important role for the derivation of climate networks. The basic mechanisms of causal discovery were presented and illustrated by examples, and available software packages from other disciplines were discussed. We believe these ideas present some exciting new research directions, but they have not yet been tested in practice. Thus the next step in our research is to apply the different algorithms for causal discovery to real-world climate data to obtain causal discovery networks. The resulting networks will be compared to each other and to correlation networks. Once we obtain reasonable network models we plan to experiment with the inclusion of selected external influences as nodes in the networks, and with varying node resolution in different geographical areas.

[1] To measure the strength of an edge between $X$ and $Y$ in an undirected graph, one could use the conditional mutual information of $X$ and $Y$ given their neighbors, i.e. $MI(X, Y|Z)$, where $Z$ is the set of all *neighbors* of $X$ and $Y$ with $X$ and $Y$ removed from $Z$.

# References


[1] B. Abramson, J. Brown, W. Edwards, M. Murphy, and R Winkler. Hailfinder: A bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12:57–71, 1996.

[2] Christian Borgelt. A conditional independence algorithm for learning undirected graphical models. *Journal of Computer and System Sciences - Special Issue on Intelligent Data Analysis*, 76(1):21–33, Feb 2010.

[3] David R. Brillinger. Second-order moments and mutual information in the analysis of time series. In Y.P. Chaubey, editor, *Recent Advances in Statistical Methods (Montreal, QC, 2001)*, pages 64–76, London, 2002. Imperial College Press.

[4] E. Burnside, D. Rubin, and R. Shachter. A bayesian network for mammography. In *Proceedings AMIA Symposium*, pages 106–110. American Medical Informatics Association, 2000.

[5] R. Cano, C. Sordo, and J.M. Gutierrez. Applications of bayesian networks in meteorology. In *Advances in Bayesian Networks*, pages 309 – 327. Springer, 2004.

[6] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence Journal*, 1-2(137):43–90, 2002.

[7] Jie Cheng, David Bell, and Weiru Lui. Learning belief networks from data: An information theory based approach. In *Proceedings of the Sixth ACM International Conference on Information and Knowledge Management (CIKM'97)*, pages 325–331, Las Vegas, NV, 1997.

[8] D.M. Chickering and C. Meek. Monotone dag faithfulness: A bad assumption, 2003. Technical Report MST-TR-2003-16, Microsoft Research.

[9] A. Cofino, R. Cano, C. Sordo, and J.M. Gutierrez. Bayesian networks for probabilistic weather prediction. In *Proceedings of the 15th European Conference on Artifical Intelligence (ECAI 2002)*, pages 695–700, 2002.

[10] G.F. Cooper and E. Herskovitz. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:330–347, 1992.

[11] M. Cossention, F.M. Raimondi, and M.C. Vitale. Bayesian models of the pm 10 atmospheric urban pollution. In *Proceedings Ninth international conference on modeling, monitoring and management of air pollution: Air Pollution IX*, pages 143–152, Ancona, Italy, September 2001.

[12] J.F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *Europhysics Letters*, 87:48007 (6pp.), 2009.

[13] I. Ebert-Uphoff. Tutorial on how to measure link strengths in discrete bayesian networks. Technical report, School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA (USA), Sept 2009. Report Number GT-ME-2009-001.

[14] N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Uncertainty in Artificial Intelligence (UAI'99)*, 1999.

[15] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3/4):601 – 620, 2000.

[16] A. Gozolchiani, K. Yamasako, O. Gazit, and S. Havlin. Pattern of climate network blinking links follws El Niño events. *Europhysics Letters*, 83(2):28005 (5pp.), July 2008.

[17] Mieczyslaw A. Klopotek. Mining bayesian network structure for large sets of variables. In *Foundations of Intelligent Systems*, volume 2366 of *Lecture Notes in Computer Science*, pages 114–122. Springer Verlag, 2002.

[18] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 1st edition, 2009.

[19] B. Lee and J. Joseph. Learning a probabilistic model of rainfall using graphical models. Project Report for Machine Learning (Fall 2006), School of Computer Science, Carnegie Mellon University. Available at http://www.cs.cmu.edu/ epxing/Class/10701-06f/project_reports.html, 2006.

[20] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *NIPS workshop on New Problems and Methods in Computational Biology*, Whistler, Canada, December 2004.

[21] Kevin P. Murphy. Active learning of causal bayes net structure. Technical report, Department of Computer Science, University of California, Berkeley, CA, 2001.

[22] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Interference*. Morgan Kaufman Publishers, San Mateo, CA, revised second printing edition, 1988.

[23] Judea Pearl. *Causality - Models, Reasoning and Inference*. Cambridge University Press, reprinted with corrections edition, 2000.

[24] G. Rebane and J Pearl. The recovery of causal poly-trees from statistical data. In *Proceedings, 3rd Workshop on Uncertainty in AI*, pages 222 – 228, Seattle, WA, 1987.

[25] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. Technical report, Philosophy, Methodology, Logic, Carnegie Mellon University, August 1990. Report CMU-PHIL-15.

[26] Peter Spirtes, Clark Glymour, and Richard Scheines. Causality from probability. In *Proceedings of the Oak Ridge Conference on Advanced Computing for the Social Sciences*, Williamsburg, VA, 1990.

[27] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

[28] Karsten Steinhaeuser, Nitesh V. Chawla, and Auroop R. Ganguly. An exploration of climate data using complex networks. In *Proceedings 3rd International Workshop on Knowledge Discovery from Sensor Data (SensorKDD'09)*, pages 23 – 31, 2009.

[29] Karsten Steinhaeuser, Nitesh V. Chawla, and Auroop R. Ganguly. Complex networks in climate science: progress, opportunities and challenges. In *Proceedings 2010 Conference on Intelligent Data Understanding*, pages 16 – 26, 2010.

[30] A.A. Tsonis and P.J. Roebber. The architecture of the climate network. *Physics A: Statistical and Theoretical Physics*, 333:497–504, February 2004.

[31] A.A. Tsonis and K.L. Swanson. Topology and predictability of El Niño and La Niña networks. *Physical Review Letters*, 100(22):228502–1–4, 2008.

[32] A.A. Tsonis, K.L. Swanson, and S. Kravtsov. A new dynamical mechanism for major climate shifts. *Physical Review Letters*, 100(22):228502–1–4, 2008.

[33] Anastasios A. Tsonis, Kyle L. Swanson, and Paul J. Roebber. What do networks have to do with climate? *Bulletin- American Meteorological Society*, 87(5):585–596, 2006.

[34] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertinaty in Artificial Intelligence*, pages 220 – 227, July 1990.

[35] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks around the globe are significantly affected by El Niño. *Physical Review Letters*, 100(2):228501–1–4, June 2008.

[36] K. Yamasaki, A. Gozolchiani, and S. Havlin. Climate networks based on phase synchronization track El-Niño. *Physical Review Letters*, 100(2):228501–1–4, June 2008.

[37] Yi-feng Zeng and Jorge Cordero Hernandez. A decomposition algorithm for learning bayesian network structures from data. In *Advances in knowledge discovery and data mining: Proceedings 12th Pacific-Asia Conference (PAKDD 2008)*, May 2008.

[38] Yi-feng Zeng and Kim-leng Poh. Block learning bayesian network structure from data. In *Proceedings Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, 2004.

[39] Yi-feng Zeng, Yanping Xiang, Jorge Cordero Hernandez, and Yujian Lin. Learning local components to understand large bayesian networks. In *2009 Ninth IEEE International Conference on Data Mining*, 2009.